Zero-Shot Object Detection

Ben Withey and Jungseok Hong

Faculty Mentor: Junaed Sattar

irvlab.cs.umn.edu

Introduction

Objective:

- Create a neural network that can detect and classify objects outside of the classes it was trained on

Motivation:

- Reduce training time since not all classes would need to be in the training set
- Allow for detections of classes that may be uncommon and/or cumbersome to collect images of

What is Zero-Shot Detection?

Main Idea:

- The concept of Zero-Shot Unseen Detection (ZSD) is being able to detect objects that weren't included in the training data

Using Wordvectors

- Instead of using visual features to make detections, this implementation ZSD maps those visual features onto wordvectors representing the objects the network was trained on to collect semantic information from the image
- The network can then use the semantic information gathered to make predictions on objects the network hasn't seen before based on a larger vocabulary matrix of wordvectors using the following equation from [1]

$$p_u = W_u W_s^T \sigma(\delta(W_s M D) f)$$

where W_u and W_s represent the wordvectors for the unseen and seen classes respectively and D is the predefined vocabulary and M is the parameters connecting seen and vocabulary wordvectors

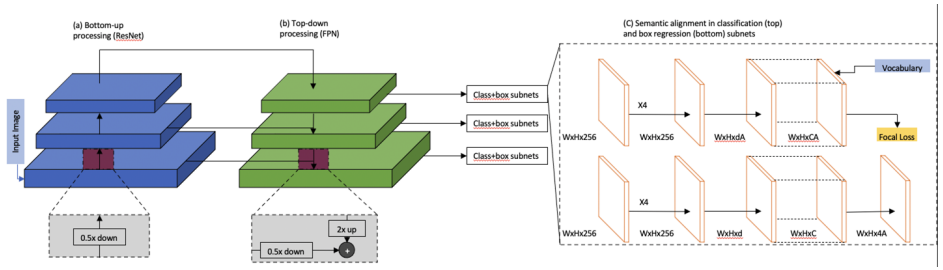

Fig. 1 A modified Retinanet network with added layers from [1] to map the physical features of an image to word embeddings

The Network

Classification Subnet:

- Adds additional layers to map physical features onto the wordvectors

Regression Subnet:

- Adds additional layers to provide the box-predictions with semantic information

Our Changes to the Network

Loss Function:

- Instead of using Polarity Loss we used Focal Loss
  - Focal Loss: Puts more weight into classifying harder examples, and reduces weight put into easy examples

Machine Learning Library

- With the original implementation being out of date it was decided to reimplement with pytorch instead of the originally used tensorflow

Training

- The network was trained using 65 of the 80 classes in the MSCOCO 2014 dataset
- The 65 training classes are the seen classes consisting of 62,3000 images
- The 15 remaining classes not included in training are the unseen classes
- The network was trained through 50 epochs

Results

Fig. 2 Seen classes are in yellow boxes, and unseen classes are in purple

Fig. 3 The training class loss

Conclusion and Work-in-progress

- This research shows that using wordvectors to detect unseen classes is feasible
- Ongoing training is expected to yield improved performance

References

1. S. Rahman, S. H. Khan and N. Barnes, "Polarity Loss for Zero-shot Object Detection," arXiv preprint arXiv:1811.08982, 2020.

Interactive Robotics and Vision Laboratory (IRVLab)
Department of Computer Science and Engineering